DOCUMENT RESUME

ED 086 728 TM 003 379

AUTHOR Hoke, Gordon A.
TITLE Evaluation and Accountability: A Resource Unit for Educators.
INSTITUTION Illinois Univ., Urbana. Center for Instructional Research and Curriculum Evaluation.
PUB DATE Jul 71
NOTE 48p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Behavioral Objectives; *Bibliographic Citations; Curriculum Evaluation; *Educational Accountability; *Evaluation; Guidelines; *Information Sources; Program Evaluation; Quality Control; *Resource Guides; Testing

ABSTRACT

This resource unit gives administrators at the elementary and secondary levels practical help in the area of assessment, evaluation and accountability. The first section deals with basic sources of information on models and conceptualizations of full program evaluations. The second section cites magazine articles and special monographs, which are shorter and more specific treatments than the basic sources. Brief descriptions of products available from agencies concerned with the evaluation of materials and instruction constitute the third section. The fourth section contains two selective bibliographies. Section five includes a variety of instruments and samples of working guidelines useful in collecting, analyzing, and interpreting evaluative data. The final section is a glossary of terms. (MP)

EVALUATION AND ACCOUNTABILITY
A RESOURCE UNIT FOR EDUCATORS

July, 1971

Issued by

Office of the Superintendent of Public Instruction

Michael J. Bakalis, Superintendent

Prepared by

The Center for Instructional Research and
Curriculum Evaluation
University of Illinois
Urbana, Illinois

in cooperation with

Division of Research, Planning and Development
Department of Research

FOREWORD

The ever-increasing demand at all levels of education for accountability to clients, taxpayers, teachers, administration, boards of education and the legislature has resulted in an emphasis on research and evaluation activities. This new perspective has caused members of the educational establishment to seek ways of answering questions about accountability based on rational and, hopefully, empirically based data.

The following unit has been developed to provide information to the educational community so that questions can be answered in a more professional and knowledgeable manner. Several source documents are provided for reference, as well as models which can be used, and a glossary of the most common terms used in connection with research, evaluation and accountability.

It is hoped that this unit will assist the members of the school community as they seek to develop a better program of instruction for the young people of Illinois. This unit is also important as a representative document developed by cooperative effort between the Office of the Superintendent of Public Instruction and another educational institution in the State.

Michael J. Bakalis
Superintendent of Public Instruction

TABLE OF CONTENTS

INTRODUCTION

Growing financial constraints and a changing social context have placed education in the public spotlight more than before. Evaluation of individual student learning, of programs, of curricula, of school processes and transactions, of goals, and of teaching behavior, is therefore urgently demanded. Students, parents, and taxpayers have always "evaluated" the schools, but seldom with much urgency and almost never systematically. In a pluralistic society, rational men can disagree, particularly about the important purposes and goals of education. This leads to a requirement of multiplicity in evaluative approaches. However, educators (teachers, supervisors, and administrators) simply must see that a large part of the acts of evaluation is kept in professional hands, and that means in their hands.

This booklet is intended to give practicing educators at the elementary and secondary levels some practical help in the assessment-evaluation-accountability job. One of the questions we hear most is: "Where can I find a source that will tell me about a method of evaluating _____?" The blank may be filled with "our curriculum" or with "my teaching" or with "student learning in my course." The pages which follow put most emphasis on answering this question, although some of the space has been used to give illustrations of useful instruments.

The first section deals with basic sources which can lead the reader to models and conceptualizations of the full evaluation scene. The second section cites magazine articles and special monographs which should be available to the school person wanting to do a systematic job. Many of these are shorter and more specific treatments than are the basic sources. Brief descriptions of products available from agencies concerned with the evaluation of materials and instruction constitute the third section. The fourth section contains two selective bibliographies, both of which are designed to complement other sections. Section five includes a variety of instruments and samples of working guidelines useful in collecting, analyzing, and interpreting evaluative data. The final section is a glossary of terms which readers may wish to glance through while studying the material in other sections. Each part of the Resource Unit is capable of standing alone and also of serving as an integral part of the total work.

Between the divisions are short "linking statements" which will help you see the connections among the different materials. You may wish to look through these linking statements before going into any one section.

Gordon A. Hoke did most of the annotations and the connecting passages with review help from others. Ludwig W. Nemeth contributed in locating sources and reviewing the work. The booklet has been tried out with some practicing school people and they have found it helpful. We hope that you do also.

PART I -- BASIC SOURCES

Source: AFT-QUEST Program. Department of Research, American Federation of Teachers, 1012 14th Street N.W., Washington, D.C. 20005

This relatively new service provides both reports and topical papers. The former are employed to publicize the results of QUEST conferences and to report on major topics of broad appeal. Papers deal with more specific issues and are prepared on a regular schedule.

Reports and papers alike range across several areas and cover issues other than those of evaluation and accountability. However, each publication presents a type of evaluative comment on different aspects of the educational scene. For example, Paper No. 12 is entitled, "The Paradigm for Accountability" and delves into the realm of teacher education. "Quality Teaching," Paper No. 3, also contains a brief statement on the evaluation of inservice training.

QUEST Reports are more comprehensive and include bibliographies. Viewpoints of teacher organizations seem certain to become more influential in the determination of school policy and practice in the forthcoming decades, a probability which adds significance to the material found in this source.

Source: "Assessment of Learning Outcomes" by J. Thomas Hastings; in The Supervisor: New Demands--New Dimensions (edited by William H. Lucio), Association for Supervision and Curricula Development, NEA, 1201 Sixteenth Street N.W., Washington, D.C. 20036, 1969. ($2.50)

Although the setting in which this presentation was made dictated that remarks be directed to school supervisors, the recent surge of interest in evaluation suggests that the article could readily serve as a general resource. The author's stress on the need to provide a broad framework for the examination of learning outcomes is the key element.

Hastings briefly reviews and analyzes the work of leading authorities in the field of evaluation and measurement. His remarks range from appraisal of the broad descriptive goals of Robert Stake to an overview of the precise behavorial objectives of Mager. The constant emphasis on the merits of full context of evaluation concludes with suggestions for the training and skills essential to thorough assessment of learning outcomes.

Readers will find the graphic illustrations in this source to be extremely valuable as they reflect the basic structure of four major approaches to evaluation. The brief bibliography is tied directly to comments in the paper and represents some of the most significant writings in evaluation of the past two decades.

At a time when pressures are mounting for schools to hastily respond to demands for accountability, this paper represents a thoughtful rejoinder that learning is, indeed, a complex matter.

Source: Educational Evaluation: Official Proceedings of a Conference, Ohio Department of Education, Columbus, Ohio 43200, 1969.

This booklet is the product of a conference funded through Title V of the Elementary and Secondary Act of 1965. Resource people included public school educators, personnel from state departments of instruction, scholars from the university community, and school board members. Consequently, readers will find that the material covers a wide range of issues and provides excellent diagnoses of important topics.

Unlike many other publications in the field, **Educational Evaluation** clearly distinguishes between research and evaluation in its opening section. Explanatory charts and diagrams accompany the more technical presentations. Only one of the nine sections has a bibliography. However, the latter appears as part of a closing presentation entitled "Current Problems in Educational Evaluation and Accountability," and adds pertinent background information to an area of growing interest and concern for both educators and laymen.

Educational Evaluation should serve schoolmen as a valuable general reference. Each of its sections represents an outstanding benchmark to guide future explorations of topics related to the broad realm of evaluation and accountability.

Source: **Handbook on Formative and Summative Evaluation of Student Learning, Benjamin S. Bloom, J. Thomas Hastings, George F. Madaus, New York: McGraw-Hill Book Company, 1971.**

The authors indicate that this book is a comprehensive report on the "state of the art" of evaluating student learning. It is intended, in their words, primarily for present and future classroom teachers. However, the work should serve as an excellent resource for all individuals and groups interested in the field of evaluation.

Although this document represents a fairly massive compendium, (it contains over 900 pages) readers will find numerous guides to enhance their reading and interpretation of contents. Part one consists of 12 chapters and an appendix. This part was prepared by the authors and provides a thorough treatment of the major substantive issues in evaluation of student learning, although it only touches upon curriculum evaluation in a few chapters. The second half is composed of a series of chapters written by scholars representing 11 areas of specialization. The topics covered in Part 2 range from preschool development through all realms of subject-matter knowledge along with an evaluation of their relationships to the instructional and learning contexts of formal education.

Diagrams, charts, and a host of operational examples highlight both parts of the work. Name and subject indices expedite usage of the material. Educators will discover that this book provides a definitive picture of the problems and possibilities suggested by the linking of evaluation to program development in our schools.

Source: **The Specification and Measurement of Learning Outcomes, David A. Payne, Waltham, Massachusetts: Blaisdell Publishing Company, 1968.**

The author cites his basic purpose as an attempt to provide classroom teachers "with a practical and efficient set of techniques to aid in evaluating student achievement." His opening chapter distinguishes between "measurement," "test," and "evaluation" and furnishes guidelines for examining the wealth of remaining material. The major focus of the book is on measurement.

Payne submits that experienced teachers should find the contents to be helpful in their daily work. His view may be somewhat optimistic since, as he acknowledges, undergraduate teacher training is frequently deficient in developing measurement competencies. Current pressures on schools to become more accountable for student performance, though, are resulting in numerous inservice training programs designed to improve teacher and administrator understanding of issues encompassing educational measurement.

The book should be a fine resource for inservice education, local study groups, etc. Its most effective use will require careful planning and skillful instructional leadership. The author provides suggested readings at the close of each of the ten chapters, supplies a comprehensive bibliography, and includes explanatory tables and charts throughout his work. Merits and demerits of various approaches and instruments are noted.

Schools interested in launching inservice education programs as a response to cries for increased accountability will find **The Specification of Learning Outcomes** a worthwhile tool.

Linking Section 1 -- General to the Specific

The five basic sources described in the preceding pages represent general references in the areas of evaluation and accountability. Educators interested in pursuing issues raised by the authors identified in Part I will wish to explore sources noted in the next few pages.

Learning is a complex matter, and the educational process cannot be honestly assessed without reliance on a variety of measures. Good men will disagree strongly about the proper approaches to use, but there is consensus that evaluation is a demanded and demanding task. Indications abound in the contemporary scene to suggest that cries for more accountability on the part of educational institutions are closely tied to fiscal affairs. True, education in a complex society is linked to economic costs as well as benefits. Yet schools also are caught up in social - political controversies. Any attempt to fairly evaluate their efforts must account for such complexity.

Materials cited in Part II provide opportunities for exploring the many facets of evaluation. Vivid illustrations of the intermingling of economic-sociological-political elements support the statements above. The works found in this section are deserving of scholarly attention, for they discuss some of the most profound problems and possibilities in American society.

PART II -- PAMPHLETS, MONOGRAPHS, AND SPECIAL MAGAZINE ISSUES

Source: "Accountability in Education," Educational Technology, January, 1971.

The entire issue of this relatively new journal is devoted to the theme of accountability. A comprehensive treatment is provided although the major emphasis rests on a systems approach encompassing a variety of works dealing with performance objectives and educational audits for both rural and urban districts.

Authors represented in this source include university and college professors and administrators; representatives of state departments of education; and public school administrators. Analysis of accountability demands in Vo-Tech Education and a description of an attempt by a state department of education to create a statewide evaluation program underscore the range of topics covered in this issue.

Copies of **Educational Technology** are available for $3.00; discounts are given for bulk orders. Reprints of individual articles can be obtained for 25 cents each. Inquiries should be directed to:

Educational Technology
140 Sylvan Avenue
Englewood Cliffs, New Jersey 07632

Source: AERA Monograph Series on Curriculum Evaluation. Rand McNally and Company, P.O. Box 7600, Chicago, Illinois 60680.

Prospective users of this series should not be misled by the title. The six issues currently available deal with instructional objectives and classroom observation, for example, in addition to an expected emphasis on curriculum.

Public school personnel are likely to find Volumes 1, 3, and 6 the most helpful. **Perspectives of Curriculum Evaluation**, the initial release, contains an article by Michael Scriven on "The Methodology of Evaluation" regarded as a landmark effort in the field. **Instructional Objectives**, the third publication, includes a presentation by W. James Popham, a pioneer in the development of behavioral objectives. Each of the four papers presented in this issue is followed by a discussion of its major points. The latest item entitled **Classroom Observation** examines the state of the art in attempts to derive certain principles of learning from better understanding of the relationship between classroom transactions and student growth.

Each of the six publications is marked by extensive bibliographies and comments by the series editor enhance understanding of the contents.

Source: Behavioral Objectives: Science, Social Studies, Mathematics, Language Arts, A Guide to Individualized Learning. John C. Flanagan, William M. Shanner, Robert F. Mager. Westinghouse Learning Corporation, 2680 Hanover Street, Palo Alto, California 94304, 1971.

The point of view that no single source of information is adequate as a basis for wise decision-making guided the preparation of this series, so say its authors.

Objectives found in the books originated from teachers and have been tried out in schools participating in Project PLAN, itself a partial outgrowth of the Project Talent study begun in 1960. Grades 1-12 are encompassed in each unit with the contents divided by levels: Primary (1-3); Intermediate (4-8); Secondary (9-12).

The books are cross-referenced and cross-indexed. There is considerable overlapping of items related to four basic subject matter areas and the vast majority of objectives are identified with the cognitive area of learning.

Source: Educational Evaluation and Decision-Making. The Eleventh Phi Delta Kappa Symposium on Educational Research. Phi Delta Kappa, Incorporated, Eighth and Union Street, Bloomington, Indiana 47401.

Phi Delta Kappa used the services and facilities of the Ohio State University Evaluation Center to hold a conference which involved a discussion of this work. Representing the efforts of a talented and diverse Study Commission, the theme of the final report reflects an approach to evaluation identified with Daniel L. Stufflebeam, Director of the Center at Ohio State and Chairman of the Phi Delta Kappa Commission on Evaluation.

Evaluation is defined as "the process of delineating, obtaining, and providing useful information for judging decision alternatives." And its purpose is seen as an attempt to improve not to prove. The reader is informed that the book is intended for a varied audience, although comments made later suggest that its most appropriate use is for evaluation units functioning inside educational institutions.

The dilemma faced by conference participants, and a problem candidly admitted by writers of the contents, concerns the complexities of the decision-making process. If evaluation provides "useful information" for those who must decide, what information is of most worth? Answers to that perplexing question are not found in the book but it represents the culmination of a time-consuming and arduous assignment by members of the Commission. The results provide innumerable ideas, suggestions, and valuable counsel.

Source: Phi Delta Kappan, December, 1970 issue

Eight articles, plus the editorial page, spotlight the theme of "accountability" in this issue. All of the authors place major emphasis on a systems approach to assessing educational outcomes. Readers will find approximately 40 pages dominated by terminology taken from the fields of economics and engineering.

Although guest editor Myron Lieberman expresses concern that the articles do not present a case for alternatives to public schools, implementation of the ideas presented in them surely would result in drastic changes. In fact, Leon Lessinger's comments pay particular attention to performance contracting by private enterprise, a phenomenon which has stirred much controversy in recent months. Concern for fiscal aspects of accountability, perhaps a logical response to legislative and taxpayer attacks, permeates these pages.

At first glance, one gains the impression that the magazine provides a comprehensive treatment of accountability, but small schools are likely to find the time and cost demands, as suggested here, overpowering. Also, little space is devoted to the problem of definition of goals.

Reprints of articles are available in minimum lots of 100. Inquiries should be directed to:

Business Office
Phi Delta Kappan
Eighth Street and Union Avenue, Box 789
Bloomington, Indiana 47401

Source: Review of Educational Research: Educational Evaluation. American Educational Research Association, 1126 16th Street, N.W., Washington, D.C. 20036, April, 1970.

Critical reviews of some of the most important work in evaluation are found in this issue. While the major focus predictably rests on a review of pertinent research in fields ranging from an analysis of measurement techniques to the assessment of social action programs, there is much valuable information here for public school officials charged with responsibility for conducting evaluation.

The material also places education in the context of public policy-debates and political exchanges concerning the school's role in a pluralistic society. Readers will discover that contributors to this issue deal with some emerging trends in evaluation and accountability, particularly the interrelationships between goals-values-national priorities and the competing claims on all social institutions, including education.

Bibliographies are very comprehensive, but the greatest value of this publication may well be its attempt to use past developments and current practices as a gauge for judging future needs in the broad panorama of educational evaluation.

Linking Section 2 -- Reliability Checks

Few institutions are ready to launch comprehensive programs of evaluation. Decisions about life in the schools cannot be taken lightly, a caution firmly underscored by the authorities noted in Part II. Nevertheless, how can educators obtain some of the most basic tools needed to carry on evaluation? Where can they find credible and helpful information?

Answers to such questions are suggested by the sources found in the upcoming section. Developmental efforts, where applicable, were handled by reputable groups; the products emanating from these activities have been field tested by practicing teachers and administrators. Training and information sites are accountable to government agencies, to private foundations, and to a broad spectrum of clients.

Not all the "answers" will be satisfactory, and some of the most perplexing questions have no ready solutions. But the aspiring evaluator should find the resources cited in Part III capable of serving as the crux of a "quality control" system for evaluating institutional policies and practices and his relationship to them.

Part III -- SPECIAL AIDS AND RESOURCES

Source: Accountability Notebook.

Gifted Children Section
Department of Exceptional Children
Office of the Superintendent of Public Instruction
Springfield, Illinois 62706

Prepared in the form of a looseleaf notebook, this resource lends an added dimension of "responsibility" to the theme suggested in the title. Accountability is regarded as including elements other than the economic one. Descriptive statements covering the broad domain of educational accountability are followed in each case by illustrations of monitoring procedures that might be employed by school systems. A third section leaves room for examples of local actions.

Source: Classroom Report.

Gifted Children Section
Department of Exceptional Children
Office of the Superintendent of Public Instruction
Springfield, Illinois 62706

This item was developed as a reporting device for teachers. It is designed to describe the class as a whole rather than individual students. Teachers can use it as one way of responding to certain calls for "Accountability"; i.e., reporting on class activities, expectations, and barriers. A manual featuring a collection of examples serves as a guide.

Source: EPIC Diversified Systems Corporation.

P.O. Box 13052
Tucson, Arizona 85711

This organization is an outgrowth of a Title III, ESEA, Project entitled Project EPIC. Originally funded as a model prototype for assisting schools with the evaluation of curricula, the enterprise has spawned training programs and the publication of evaluative materials. Major efforts are focused on three areas: 1) Accountability services, including educational audits; 2) Comprehensive planning and evaluation with an emphasis on needs assessment; 3) Leadership training institutes.

Source: Instructional Objectives Exchange.

Center for the Study of Evaluation,
University of California
Los Angeles, California 90024

The Exchange is devoted to the collection and distribution of operationally stated instructional objectives and related evaluation measures. Sets of objectives and items in virtually all subject areas are available beginning at the kindergarten level. Educators can select materials from the depository which appear most suitable to their specific requirements.

Source: The Educational Product Report.

Educational Products Information Exchange
(EPIE) Institute, 386 Park Avenue South
New York, New York 10016

EPIE publishes this item as a forum for providing descriptive and evaluative information and commentary about all types of learning materials, equipment, and systems. The Report is supported entirely by subscriptions and does not accept advertising. Interested persons will find the February, 1969, issue on "Educational Evaluation: Theory and Practice" particularly useful.

Linking Section 3 -- Other References

A major barrier to increased use of evaluation is the narrow image many public school teachers and administrators have of the field. Rarely has their undergraduate or graduate training provided sustained contacts with the questions of assessment and decision-making now confronting them. The preceding pages represented an effort to pinpoint specific sources of help. Part IV builds on that base, extends it, and attempts to highlight some of the latest ideas and approaches to evaluation.

References listed in Item A are cited in "Assessment of Learning Outcomes," one of the five Basic Sources found in Part I. Hastings refers to them in the context of his paper, furnishing operational examples and providing linkage between various ideas contained in these important works.

Item B is comprised of ten references designed to complement the areas and issues covered in other sections of this Resource Unit. In some cases, the listing refers to another bibliography. Readers will note the diverse range of topics covered here, for the list reflects the rich ferment now marking developments in evaluation and accountability.

PART IV -- RELATED BIBLIOGRAPHIES

(A)

1) Atkin, J.M., "Some Evaluation Problems in a Course Content Improvement Project," **Journal of Research in Science Teaching**, 1,129-132, 1963.

2) Bloom, [illegible]in S., ed., **Taxonomy of Educational Objectives, Handbook 1; Cognitive Domain**. New York: Longmans, Green and Co., 1956.

3) Cronbach, L.J., "Course Improvement Through Evaluation," **Teachers College Record**, 64, 672-683, 1963.

4) Dressel, Paul L., and Lewis B. Mayhew, **General Education: Explorations in Evaluation.** Washington, D.C.: American Council on Education, 1954.

5) Easley, John A., Jr., "The General and the Particular in Educational Research," **Curriculum Laboratory Working Paper, No. 10.** Urbana, Illinois: College of Education, University of Illinois, June 1967.

6) Gagne, Robert M., "Curriculum Research and Promotion of Learning," **AERA Monograph Series on Curriculum Evaluation, Vol. 1.** Chicago: Rand McNally and Company, 1967.

7) Gallagher, J.J., "Teacher Variation in Concept Presentation," **Newsletter of the Biological Sciences Curriculum Study**, No. 30. Boulder, Colorado: University of Colorado, January 1967.

8) Hastings, J.T., "Curriculum Evaluation: The Whys of the Outcomes," **Journal of Educational Measurement**, 3,27-32, 1966.

9) Krathwohl, David R., Benjamin S. Bloom, and Bertram B. Masia, **Taxonomy of Educational Objectives, Handbook II: Affective Domain**. New York: David McKay Company, 1964.

10) Lortie, Dan, **Rational Decisions on School Curricula, Curriculum Laboratory Working Paper, No. 5.** Urbana, Illinois: College of Education, University of Illinois, October 1966.

11) Mager, Robert F., **Preparing Objectives for Programmed Instruction.** San Francisco: Fearon Publishers, 1961.

12) Maguire, T.O., **Value Components of Teachers' Judgments of Educational Objectives**. Unpublished doctoral dissertation. Urbana, Illinois: University of Illinois, 1967.

13) Paden, D.W., "Instructional Television and the Programming Process," **National Society for Programmed Instructional Journal**, 6, 4-9, 1967.

14) Scriven, Michael, "The Methodology of Evaluation," **AERA Monograph Series on Curriculum Evaluation, Vol. 1.** Chicago: Rand McNally and Company, 1967.

15) Smith, Eugene R., and Ralph W. Tyler, **Appraising and Recording Student Progress**. New York: Harper and Brothers, 1942.

16) Stake, Robert E., "The Countenance of Educational Evaluation," **Teachers College Record**, 68,523-540, April 1967.

17) Taylor, P.A., **The Mapping of Concepts**. Unpublished doctoral dissertation. Urbana, Illinois: University of Illinois, 1966.

18) Taylor, P.A., and T.O. Maguire, "A Theoretical Evaluation Model," **The Manitoba Journal of Educational Research**, 1, 12-17, 1966.

19) Tyler, R.W., "Constructing Achievement Tests," Reprints from the **Educational Research Bulletin**, Ohio State University. Columbus: Ohio State University, 1934.

20) Webb, Eugene J., et al, **Unobtrusive Measures: Nonreactive Research in the Social Sciences**. Chicago: Rand McNally and Company, 1966.

(B)

1) Coiler, Alan R., An **Annotated Bibliography of Self-Concept Measures for the Early Childhood Years.** Urbana, Illinois: ERIC Clearinghouse on Early Childhood Education, University of Illinois at Urbana-Champaign.

2) Ebel, Robert L., "Behavioral Objectives: A Close Look," **Phi Delta Kappan,** November, 1970, pp. 171-173.

3) Eidell, Terry, and John A. Klebe, **Annotated Bibliography on the Evaluation of Educational Programs.** Eugene, Oregon: ERIC Clearinghouse on Educational Administration, University of Oregon, 1968.

4) Glass, Gene V., **The Growth of Evaluation Methodology.** Boulder, Colorado: Laboratory of Educational Research, University of Colorado.

5) Grotelueschen, Arden D., and Dennis D. Gooler, **The Role of Evaluation in Planning Educational Programs.** Urbana-Champaign, Illinois: University of Illinois, and Syracuse, New York: Syracuse University, 1970.

6) Guba, Egon, "The Failure of Educational Evaluation," **Educational Technology,** May, 1969, pp. 29-38.

7) House, Ernest R., **The Conscience of Educational Evaluation.** Urbana-Champaign, Illinois: Center for Instructional Research and Curriculum Evaluation, College of Education, University of Illinois.

8) McCall, George J., and J.L. Simmons, ed., **Issues in Participant Observation: A Text and Reader.** Reading, Mass.: Addison Wesley Publishing Co., 1969.

9) Owens, Thomas R., **Application of Adversary Proceedings to Educational Evaluation and Decision-Making.** San Jose, California: Center for Planning and Evaluation, 1971.

10) Wardrop, James L., **Determining 'Most Probable Causes': A Call for Re-Examining Evaluation Methodology.** Urbana, Illinois: Center for Instructional Research and Curriculum Evaluation, University of Illinois.

Linking Section 4 -- Working Guides

The twelve items found in Part V are labeled "The Daily Dozen" in reference to their emphasis on utility. Their selection and arrangement draw upon the results of field studies conducted in 1968-69 and also reflect judgments obtained from individuals and groups concerning the contents of this Resource Unit. There is considerable variance among these resources for they are based on reported needs in the realm of operational evaluation, i.e., meeting the functional demands of school personnel.

An arbitrary arrangement featuring four (4) clusters of task-related items characterizes this section. An introductory note precedes each division. The first division contains materials generally identified with goal selection and choice of objectives. Cluster 2 holds various types of checklists, rating forms, testing procedures, grading plans. It is followed by two outlines for evaluating total school programs. Attitudes, judgments, and opinions receive due attention in the final cluster. A sophisticated analysis of weaknesses in performance contracting is accompanied by an explanatory note written in "layman's language." Part V concludes with a scale for determining attitudes toward educational evaluation. It, too, features side comments.

PART V -- THE DAILY DOZEN

Introductory Note--Cluster 1: "Goals and Objectives"

Schools currently pursue many more objectives than any educator can specify, or that any evaluation plan can accommodate. As recent history shows, public discussion of educational goals frequently evokes strong political and social reaction. Cultural diversity has enriched our nation, and future planning based on attempts to provide means for making educated choices appears to be a necessity, both for maintaining strong school systems and for the welfare of our diverse peoples.

The two items found in this cluster point to certain procedures for consideration in choosing goals and pursuing objectives. They also emphasize the vital importance of the process used in those activities.

An important task in any evaluation is the examination of

Educational Goals

Who has goals?

Many groups of people have ideas concerning the proper goals of public education. The educational goals which a particular school system pursues reflect the ideas and influence of groups such as the administrators, teachers, boards of education, parents, religious leaders, businessmen, and professors of education.

Goals compatible?

We must often determine if goals advocated by different groups are compatible. When differences in objectives exist (as they frequently do), the evaluator may be asked to help describe these differences; for instance, to state which groups are taking which positions or goals. One of legitimate evaluation's responsibilities is the collection of information about the goal preferences of different groups.

What goals to evaluate?

Key questions are whose goals? or what goals? The question of goal priorities always emerges. The school coach, the art teacher, and a local businessman will likely differ on what educational goals they think are important.

Establishing priorities.

Evaluators should be interested in how priorities are established and may include a description of this process in their evaluation report. Questions of concern to the evaluator are: Who has the legal authority to establish priorities about educational goals? Who has the informal power to do so? Who actually makes what decisions about which goals?

Identifying (measuring) goals.

Once goals are known, the evaluator must identify or describe these goals accurately in greater detail. Controversy centers on the question of the importance of translating all educational goals into strictly behavioral terms to measure achievement or nonachievement. Evaluators are often better able to make this translation than are curriculum specialists and teachers.

Goals achieved?

How to determine whether the goals or "intended outcomes" of a program or course have been achieved? Recent writers on evaluation argue that evaluation should also provide information useful in learning "why" goals were or were not achieved. Techniques for measuring outcomes range from psychometric tests to observation schedules to anthropological studies.

For further ideas on educational goals as they relate to evaluation see:

1. Atkin, J. M., Behavioral Objectives in Curriculum Design: A Cautionary Note. **Science Teacher**, 35, No. 5, May 1968, 27-30.

2. Popham, J.W., Probing the Validity of Arguments Against Behavioral Goals. Paper read at American Educational Research Association, Chicago, February 1968.

3. Stake, R. E., The Countenance of Educational Evaluation. **Teachers College Record**, 1967, 68, 523-540.

EDUCATIONAL OBJECTIVES

1. In any teaching a great number of objectives are simultaneously pursued. High-priority, immediate objectives should usually be apparent to teacher and learner alike. Occasionally, either will do better **without** being aware of them. High-quality education is often accomplished by educators having but a partial awareness of the objectives. Sometimes it will increase teaching-learning effectiveness to make participants more aware of objectives; sometimes it will not.

2. With all who share the responsibility of educating lies the responsibility for stating objectives, arranging environments, providing stimulation, evoking responses, and evaluating those responses. But each author and teacher does not share equally in those responsibilities. Time and talent are not available in limitless abundance. Each educator's assignment should capitalize on what he can do best. Few classroom teachers are skilled in stating objectives. Most are more highly skilled in adapting teaching to immediate circumstances, motivating students, and appraising responses. In the interests of effectiveness, seldom should they be required to formulate behavioral specifications.

3. There are more objectives to pursue than we can pursue. Time and resources restrict us. We assign priorities to our goals in a highly informal way. Even this informal priority list is not always the critical determinant of the daily lesson or the minute-by-minute dialogue. Some moments are ripe for teaching toward an unplanned objective. A sound educational system is one which provides for occasional reassignment of immediate objectives to take advantage of the special opportunities that occur.

4. The development of a new curricular program or set of instructional materials often proceeds better by successive approximations than by linear programming. With successive approximations, major attention is given to getting an enterprise in operation, even though the initial runs are crude and faulty, so that corrections can be based on experience. With linear programming, major attention is given to planning, precise specification, and symbolic representation so that corrections can be based on logical analysis. Advice on curriculum planning should be oriented to the experiential and logical skills already developed in the developers or that can be readily obtained by them.

5. For creating lists of objectives, the technology of education should have some methods that rely on behavioral specification and symbolic delimitation and other methods that rely on illustrative examples and inferable definitions. We need methods by which educators and others can endorse, reject, or revise statements of objectives. Two colossal problems lie before us: how to **translate** global objectives into specific behavioral objectives and how to **derive** appropriate teaching tactics.

6. Our curriculum-development projects and our evaluation studies seldom reach a satisfactory specification by asking educators to state their objectives. Educator's global objectives give little guidance to teaching and evaluation. Their specific objectives ignore vast concerns that they have. In our present state the derivation of the specific from the general is some form of intuitive magic. Luckily it often works pretty well. We need to understand it, to simulate it, not necessarily to replace it.

Introductory Note--Cluster 2: "Checklists, Rating Forms, Testing Procedures, Grading Plans"

Respondents in the field testing of an **Evaluation Kit** in 1968-69 stressed a need for "How to" materials. The six items found in this section are a response to their comments. Readers will discover that the materials encompass varying levels of utility depending on the local situation.

Illustrations of curriculum evaluation, item sampling, and textbook analysis may be more appropriate for system-wide committee usage. Likewise, teachers will find the grading plan provocative, and administrators may wish to discuss implications of the issues covered by evaluation reports.

FORMAT FOR AN EVALUATION REPORT FOR AN EDUCATIONAL PROGRAM

SECTION I -- OBJECTIVES OF THE EVALUATION

- A. Audiences to be Served by the Evaluation
- B. Decisions about the Program, Anticipated

SECTION II -- SPECIFICATIONS OF THE PROGRAM

- A. Educational Philosophy Behind the Program
- B. Subject Matter
- C. Learning Objectives, Staff Aims
- D. Instructional Procedures, Tactics, Media
- E. Students
- F. Instructional and Community Setting
- G. Standards, Bases for Judging Quality

SECTION III -- PROGRAM OUTCOMES

- A. Opportunities, Experiences Provided
- B. Student Gains and Losses
- C. Side Effects and Bonuses
- D. Costs

SECTION IV -- RELATIONSHIPS AND INDICATORS

- A. Congruence
- B. Contingencies
- C. Trend Lines, Indicators, Ratios

SECTION V -- JUDGMENTS OF WORTH

- A. Value of Outcomes
- B. Relevance of Objectives to Needs
- C. Usefulness of Evaluation Information Gathered

A CHECKLIST FOR RATING AN EVALUATION REPORT

This checklist can be used to examine the report of an evaluation of an educational program to see if the report provides complete and useful information.

	Well Stated	Needs Better Statement	Not Stated	Not Applicable
Area I -- THE EVALUATION ITSELF				
A. Audiences to be served by the evaluation				
B. Decisions about the program, anticipated				
C. Rationale, constraints, bias of evaluators				
Area II -- SPECIFICATIONS OF THE PROGRAM BEING EVALUATED				
A. Educational philosophy behind the program				
B. Subject matter to be taught				
C. Learning objectives, staff aims				
D. Instructional procedures, tactics, media				
E. Students: biography, readiness, goals, etc.				
F. Instructional and community setting				
G. Standards, bases for judging quality				
Area III -- PROGRAM OUTCOMES				
A. Opportunities, experiences provided				
B. Student gains and losses				
C. Side effects and unexpected bonuses				
D. Costs: cash, resources, work, morale				
Area IV -- RELATIONSHIPS AND INDICATORS				
A. Congruence between intent and actuality				
B. Contingencies, causes and effects				
C. Trend lines, indicators, comparisons				
Area V -- JUDGMENTS OF WORTH OF THE PROGRAM				
A. Value of outcomes, different points of view				
B. Relevance of objectives to needs				

Readability of report _____

Usefulness of evaluation information gathered _____

Comments:

PROTOTYPES OF CURRICULUM EVALUATION

PROTOTYPE	KEY EMPHASIS	PURPOSE	KEY ACTIVITIES	KEY VIEWPOINT USED TO DELIMIT STUDY	OUTSIDE EXPERTS NEEDED	EXPECTED TEACHING STAFF INVOLVEMENT	RISKS	PAYOFF
Ralph Tyler's Evaluation Model	Instructional objectives	To measure student progress toward objectives	Specify objectives; measure student competence	Curriculum supervisor; teacher	Objectives specifiers; measurement specialists	Conceptualize objectives; give tests	Oversimplify school aims; ignore processes	Ascertain student progress
School Accreditation Model	Staff self-study	To review content and procedures of instruction	Discuss program; make professional judgments	Classroom teacher; administrator	None, unless authentication by outside peers needed	Committee discussions	Exhaust staff; ignore values of outsiders	Increase staff leadership responsibility
Bob Stake's Countenance Model	Description and judgment data	To report the ways different people see the curriculum	Discover what audience wants to know about; observe; gather opinions	Audience of final report	Journalists; social psychologists	Keep logs; give opinions	Stir up value conflicts; ignore causes	Broad picture of curriculum and conflicting expectations
Dan Stufflebeam's CIPP Model	Decision-making	To facilitate rational and continuing decision-making	Identify upcoming alternatives; study implications; set up quality-control	Administrator; director	Operations analysts	Anticipate decisions, contingencies	Overvalue efficiency; undervalue student aims	Curriculum sensitive to feedback
Hilda Taba's Social Studies Evaluation Model	Cause and effect relationships	To seek simple but enduring explanation of what works	Exercise experimental control and systematic variation	Theorist; researcher	Research designer; statistical analysts	Tolerate experimental constraints	Artificiality; ignore personal values	Get rules for developing new programs

REFERENCES:

TYLER, RALPH W. General Statement on Evaluation. *Journal of Educational Research*, March, 1942, 492-501.

NATIONAL STUDY OF SECONDARY SCHOOL EVALUATION. *Evaluative Criteria, 1960 Edition*, National Study of Secondary School Evaluation, Washington, D. C.

STAKE, ROBERT E. The Countenance of Educational Evaluation. *Teachers College Record*, 68, 1967, 523-540.

STUFFLEBEAM, DANIEL L. Evaluation as Enlightenment for Decision Making. In W. H. Beatty (Ed.), *Improving Educational Assessment and An Inventory of Measures of Affective Behavior*. Washington: Association for Supervision and Curriculum Development, NEA, 1969. Pp. 41-73.

TABA, HILDA. *Teaching Strategies and Cognitive Functioning in Elementary School Children*. Cooperative Research Project No. 2404. San Francisco State College, San Francisco, 1966.

Prepared by R. E. Stake, CIRCE, October 1969

AROO GRADING PLAN A school-wide plan for assignment of final course grades to students. Four grades are submitted for each student except where instructor feels he does not have an adequate basis for grading.

	INFORMATION	SOURCE OF OBJECTIVES	STANDARD
ABSOLUTE JUDGMENT	The individual student's quality of work and readiness to perform with competence	The goals set by the academic department	Quality and competence as valued by the instructor
RELATIVE STANDING	The individual student's standing among peers as to work quality and competence	The goals set by the academic department	Empirically determined quality and competence of the designated reference group
OPPORTUNITY USED	The individual student's effort and success in using this learning opportunity	The goals set by the individual student	Judgment by the instructor as to whether student successfully pursued any educational objectives
OPPORTUNITY PROVIDED	Quality of learning opportunity provided by school, instructor and other students	The collective goals of the students (not specified)	Judgment by the students (one mark, the same for the whole class)

Categories for reporting and interpreting grades:

Absolute Judgment:	5 = Excellent	4 = Good	3 = Fair	2 = Poor	1 = Very Poor
Relative Standing:	5 = Superior	4 = Above average	3 = Average	2 = Below average	1 = Inferior
Opportunity Used:	5 = Excellent	4 = Good	3 = Fair	2 = Poor	1 = Very Poor
Opportunity Prov.:	5 = Excellent	4 = Good	3 = Fair	2 = Poor	1 = Very Poor

From: "Grading Students in the Real World,"
Robert E. Stake (Las Cruces, Jan. 11, 1971)
Robert E. Stake, CIRCE, University of Illinois

Background reference:
Warren, Jonathan R., College grading practices: an overview. Berkeley: Educational Testing Service, 1970

One way of measuring the achievement of student groups is to use

ITEM SAMPLING

This sheet tells the story of an evaluation of Spanish instruction via television--using item-sampling techniques. To get that story, the evaluator needed a good picture of learning. During the year he tested the students for 15 minutes every two weeks on their understanding of Spanish vocabulary.

To start out, he made a pool of 360 test items, then randomly sorted them into nine tests of 40 items each. The test items looked like those at the right although most were harder.

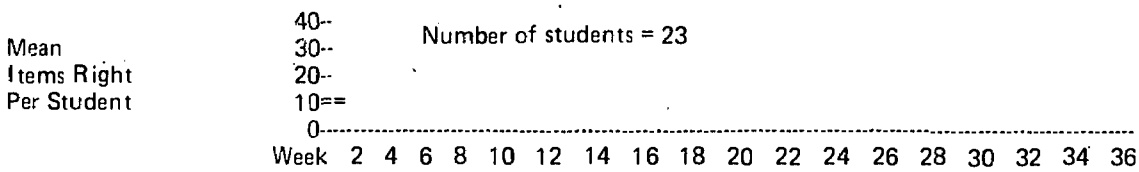
Roma es la capital de
(1) Francia
(2) Brasil
(3) Mexico
(4) el Canada
(5) Italia

Sombrero
(1) darkness
(2) sleep
(3) hat
(4) summer
(5) cow

Every second Thursday tests were randomly assigned to students. Therefore, all 360 were used twice a month even though each student was answering only 40 items twice a month. Note that the 360 items covered a lot of vocabulary each month even though student's testing time was only 30 minutes.

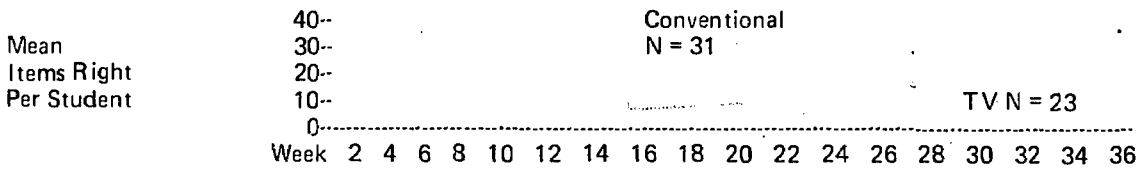
A student could draw the same test both times during the month but that did not happen often. It would have been better to have let the students take each test only once every 18 weeks but that would have required a little more work and class time.

Since the same 360-item "test" was given to the TV-taught students each testing time, group means were used to show progress in learning vocabulary across the year. Here is the curve of progress for the TV students:



This evaluation design is like a pretest-posttest design plus a lot of intermediate testing. The curve of progress tells much more than that there is a mean gain of 4 items in 30 weeks.

The results of the evaluation study are more meaningful when the evaluator shows the progress of students taught in other classes.



Of course, none of this is meaningful without an understanding of what the TV and conventional instruction consisted of. Furthermore, there are other aspects of achievement beside vocabulary, e.g., pronunciation, grammar, idiom, word derivation, literary usage, appreciation, etc.

Item sampling is a valuable evaluation tactic when the interest is in group achievement rather than individual achievement, when the content to be covered is broad, and when only a little time for testing is available.

Do you have the job of

Selecting a textbook?

A potential textbook selector who has answered the following questions (as well as others) should be able to make realistic and knowledgeable judgments concerning the selection of texts and other educational materials.

1. Why is the present text inadequate?
2. Will the adoption of this new text further the educational interests of the community?
3. What are the specific educational objectives in this content area, and how well will this text help in reaching such objectives?
4. Will this text prove compatible with present instructional methodology?
5. Will the students find the text easy to follow and comprehend? attractive?
6. Has the selection of this text been preceded by an objective consideration of other available textbooks?
7. Have appropriate book reviews, reports, or institutional comments on the text's usefulness been consulted?*
8. What provisions have been made for an ongoing evaluation of the text if it is accepted?
9. Does the publisher provide additional materials such as a teacher's handbook, workbooks, examinations, etc., to go along with the text?
10. Is the purchase of a text the best use of limited financial resources?
11. Have alternative instructional materials been investigated?

* A further discussion of problems involved in the evaluation of instructional materials can be found in:

The EPIE Forum I, No. 4-5, December 1967 and January 1968

Introductory Note--Cluster 3: "Comprehensive Evaluation"

While it is unlikely that any plan of evaluation could meet every single demand for accountability, the illustrations to follow are indicative of the countless responsibilities confronting schools and their communities.

Attempts to deal with accountability on a wide front cannot escape the implications of sharing power. When individuals and groups are included in the decision-making process, they are more willing to assume responsibility for consequent actions. Accountability carries with it definite overtones of responsibility. Unless the pivotal role of schools in modern society is understood by all of us there is real danger that only individuals, most of them involved in the instructional process, will face the consequences of our present desire to evaluate the schools.

EVALUATION REPORTS
DESCRIBING THE CONTEXT OF A PROGRAM

City or Community Characteristics

What is the population of the city or community?
What adjective(s) would typically be used to describe the city or community?
In what part of the country is it located?
What is the percentage of deteriorating or dilapidated housing in the city or community?
What is the city- or community-wide unemployment rate?
What percent of families in the city or community are on welfare?
What is the city- or community-wide literacy rate?
What is the city- or community-wide school dropout rate?
What is the city- or community-wide delinquency rate?
Are there any special educational problems faced by the city or community?
What attempts, if any, are being made to deal with these problems?

Neighborhood Characteristics

What adjective(s) would typically be used to describe the neighborhood(s)?
What is the average family income in the neighborhood(s)?
What is the literacy rate in the neighborhood(s)?
What kinds of occupations do most of the people in the neighborhood(s) have?
What is the unemployment rate of the neighborhood(s)?
What percent of the families in the neighborhood(s) are on welfare?
What is the percent of nonintact families in the neighborhood(s)?
What ethnic groups, in what percent, are represented in the neighborhood(s)?
What linguistic groups, in what percent, are represented in the neighborhood(s)?
What is the population density (number of people per square mile) in the neighborhood(s)?
What is the percent of multi-family dwellings in the neighborhood(s)?
What percent of the dwellings were built pre-1940 in the neighborhood(s)?
What percent of the dwellings are rental (rather than owner-occupied) in the neighborhood(s)?
What is the percent of deteriorating or dilapidated housing in the neighborhood(s)?
What is the school dropout rate in the neighborhood(s)?
What is the delinquency rate in the neighborhood(s)?
Have these neighborhood characteristics remained constant in the last few years or is the neighborhood(s) in transition?

School Characteristics - General

What was the per capita expenditure, including both capital and operating expenses, prior to the program?
What was the salary range for teachers in the school(s) for the year immediately preceding the program?
What is the age and condition of the main school building(s)?
What grade levels were included in the school(s)?
What was the average teacher-pupil ratio in the school(s)?
How were the students routinely grouped in the school(s)?
Were any pupils enrolled in the school(s) as a result of a bussing or open enrollment program?
Was a conventional curriculum followed in the school(s)?
What services, personnel, or special programs were available in the school(s) prior to the program?
Were any other specially funded programs ongoing in the school(s) prior to the beginning of this program?
At what intervals are achievement tests routinely given?
What achievement tests are routinely given? To what grades?
How are these achievement tests administered and by whom?
How did the achievement level of the school(s) compare with city-wide and/or national norms prior to the program?

School Characteristics - Teachers

What were the paper qualifications of the teachers?
What was the average number of years of teaching experience?
What was the average age of the teachers?
What was the male-female ratio of teachers?
What ethnic groups, in what percent, were represented by the teachers?
What linguistic groups, in what percent, were represented by the teachers?
What was the teacher turnover in the school(s) prior to the beginning of the program?

Prepared by C. O. Neidt, Colorado State University, 1969.

School Characteristics - Student Body

What was the pupil enrollment in the school(s) at the beginning of the academic year?
How many pupils withdrew or transferred from the school(s) after the school year began?
How many pupils enrolled in the school(s) after the school year began?
What was the average daily attendance in the school(s)?
Has the total pupil enrollment in the school(s) involved in the program changed in the last three years?
What ethnic groups, in what percent, were represented by the students?
What linguistic groups, in what percent, were represented by the students?
What was the male-female ratio of the students?

Historical Background

Did the program exist prior to the time period covered in the present report?
Is the program a modification of a previously existing program?
How did the program originate?
What special efforts were made to gain acceptance of the program by parents and the community before it began?
If special problems were encountered in gaining acceptance of the program by parents and the community, how were these solved so that the program could be introduced?

DESCRIBING THE TREATMENT PROVIDED BY A PROGRAM

Personnel: Instructional and Noninstructional

What categories of personnel were added by program?
What regular staff were assigned to program?
What new staff were hired for program?
What were paper qualifications for various personnel?
What were average years of relevant experience of personnel?
What were the most important duties of personnel?
What was the time commitment of various personnel?
What inservice training was provided?
What was the male-female ratio of classroom personnel?
What personnel characteristics enhanced or reduced program effectiveness?
How did special needs of pupils affect staff development and utilization?

Supporting Services

What services were part of the program?
What services were available to experimentals? To controls? To both?
How did special needs of pupils affect provision of services?

Organization: Schedules

For how long did the program operate?
How were experimental and control classes scheduled in the total school context?
How many hours of instruction did experimentals receive? Controls?
Were time intervals between learning and testing equivalent for these groups?

Organization: Planning

Were meetings held regularly for experimental and control teachers?
What were the purposes of these meetings?
Who was present (besides teachers) and why?

Organization: Physical Arrangements

Where were experimental classes located?
Where were control classes located?
What were the most noteworthy features of physical arrangements in each?

Organization: Grouping of Teachers

How were experimental and control teachers grouped for instructional purposes?

Organization: Grouping of Pupils

How were pupils grouped within the total school context?
How were pupils grouped for instruction in experimental and control classes?
How many children were in each experimental class? In each control class?

Major Program Segments

What major segments comprised program?
Which of these were available to experimentals? To controls? To both?
Were segments equivalent for these groups in the following respects:
- Objectives?
- Emphasis?
- Provision for motivating pupils?

How did special needs of pupils affect content of major program segments?
What characteristics of these segments enhanced or reduced program effectiveness?

Methodology: Pupil Activities

What were main activities of experimentals? Of controls?
How much time was devoted to each main activity?
How many pupils were involved in each?
How were instructional materials used by pupils in each?
Did pupils have freedom of choice in participating in each main activity?
How much time did pupils spend in the program each day? Each week?

Methodology: Teacher Activities

What were main activities of teachers in experimental and control classes?
How much time did the teacher spend with the pupils?
What was the teacher-pupil ratio (or aide- or adult-pupil ratio)?
What provision did the teacher make for pupil response?
How did the teacher use various instructional materials for the activity?
What provision did the teacher make for pupil response?
To what extent were teachers free to experiment with teaching methods?
How did the teacher give feedback to pupils on individual progress?
What provision did the teacher make for motivating pupils?
Were amounts of practice, review, and quiz activities equivalent for both groups?
Was content of these activities equivalent for both groups?
How did special needs of pupils affect teaching methods?
What characteristics of activities enhanced program success?

Instructional Equipment and Materials

What equipment and materials were used by experimentals? Controls? Both?
In what amounts?
What equipment and materials were used in each main activity in the two groups?
What specific features suited a given device to a particular activity?
Were materials equivalent for both groups in the following respects:
- Subject-matter content?
- Content of drill?
- Vocabulary level?

What instructional materials were developed for program? How were they developed?
What characteristics of materials enhanced or reduced program effectiveness?
How did special needs of pupils affect selection or development of materials?

Parent-Community Involvement

What provisions were made for parent and/or community involvement in the program?
Were these provisions equivalent for parents of experimentals and controls?
Were group meetings and/or parent conferences held for parents of experimentals and controls? Describe.

Budget

What was the total cost of program? (indicate length of time covered)
From what sources were these funds obtained?
What portion of total program cost was start-up expense? Continuation expense?
Can you break down total program cost into broad categories of expenses?
If the program were repeated, how would you modify the budget?
What was per-pupil cost of program?
How does it compare with normal per-pupil cost of schools in the program?
Where can the reader get additional budget information?

DESCRIBING, ANALYZING AND INTERPRETING EVIDENCE OF CHANGES INDUCED BY A PROGRAM

Objectives:

What was the program aiming to do for the children and adults in it?

Were the children expected to improve their scores on achievement measures? If so, in what areas?
Were the teachers or other adults expected to change their modes of instruction?
Were the children expected to change their attitudes? If so, which ones?
Were the teachers or other adults expected to change their attitudes? If so, which ones?

Sampling Procedures:

How were the children and adults in the program chosen?

Were the samples originally representative of the populations from which they were chosen?
Were the controls selected before or after the program?
Were steps taken to avoid the samples being affected by other programs?
Were steps taken to avoid real differences in the quality of teachers selected for experimental and control groups?
Was there attrition of the samples?
Was there attrition of groups of children with the same characteristics?
Were pupils added to the samples to replace dropouts?
Were there many children who did not receive the treatment often because of poor attendance?
Did the children participate voluntarily?
Were the same children included in both pretest and posttest samples?

Describing Samples:

Which children received the treatment, from which adults?

What is the size of the experimental sample?
What is the age or grade level of the experimental sample?
How is the experimental sample divided into boys and girls?
Are achievement scores available by which to describe the experimental sample?
Which adults gave the treatment that constituted the program?

Measuring Change:

What measures were applied to find out whether the program's aims had been achieved?

Were the measures matched to the objectives in content?
Did the tests used have sufficient "floor" and "ceiling"?
Were the same measures used for both experimental and control groups?
Were the same measures (or parallel forms) used for both pre and posttesting?
Were IQ tests used when achievement tests were more appropriate?
Was the reliability of the tests quoted?
Under what conditions were the measures applied?
Were the same or different testers used for successive testings?
Were oral, or written, instructions available for the tests?
Were assessors or observers likely to bias the results for or against the program?
How much time elapsed between testings?
Were assessors or observers specially trained?

Presenting Data:

What data were obtained from the measures applied?

What measures of central tendency should be used?
What measures of dispersion were used?
Were there graphical displays which could have been used to present data more clearly?

Analyzing Data:

What analyses were undertaken of the data?

Was there a proper basis against which to compare the progress of the experimental group?
What was the correlation between pretest and posttest?
What comparisons were drawn for subsamples?
Is there any evidence that children who attended more gained more from the program?
Was the formula or source given for the statistical test applied?
Did the date meet the prerequisites for the statistical tests used?
Were there real differences between the groups?

Drawing Conclusions:

What conclusions were drawn from the analyses of the results?

Were the conclusions based on statisticalprobability?
Were the statistical conclusions translated into ordinary language?
Were other conclusions stated in ordinary language?
Can the conclusions be generalized, or are they applicable only to the sample or population served by the program?
Were the conclusions of educational importance?
What recommendations can be based upon the conclusions?

MULTIPLE CRITERION MEASURES FOR EVALUATION OF SCHOOL PROGRAMS*

Newton S. Metfessel and William B. Michael
University of Southern California

I. Indicators of Status or Change in Cognitive and Affective Behaviors of Students in Terms of Standardized Measures and Scales.

Standardized achievement and ability tests, the scores on which allow inferences to be made regarding the extent to which cognitive objectives concerned with knowledge, comprehension, understandings, skills, and applications have been attained.

Standardized self inventories designed to yield measures of adjustment, appreciations, attitudes, interests, and temperament from which inferences can be formulated concerning the possession of psychological traits (such as defensiveness, rigidity, aggressiveness, cooperativeness, hostility, and anxiety).

Standardized rating scales and check lists for judging the quality of products in visual arts, crafts, shop activities, penmanship, creative writing, exhibits for competitive events, cooking, typing, letter writing, fashion design, and other activities.

Standardized tests of psychomotor skills and physical fitness.

II. Indicators of Status or Change in Cognitive and Affective Behaviors of Students by Informal or Semiformal Teacher-made Instruments or Devices.

Incomplete sentence technique: categorization of types of responses, enumeration of their frequencies, or ratings of their psychological appropriateness relative to specific criteria.

Interviews: frequencies and measurable levels of responses to formal and informal questions raised in a face-to-face interrogation.

Peer nominations: frequencies of selection or of assignment to leadership roles for which the sociogram technique may be particularly suitable.

Questionnaires: frequencies of responses to items in an objective format and numbers of responses to categorized dimensions developed from the content analysis of responses to open-ended questions.

Self-concept perceptions: measures of current status and indices of congruence between real self and ideal self -- often determined from use of the semantic differential or Q-sort techniques.

Self-evaluation measures: student's own reports on his perceived or desired level of achievement, on his perceptions of his personal and social adjustment, and on his future academic and vocational plans.

Teacher-devised projective devices such as casting characters in the class play, role playing, and picture interpretation based on an informal scoring model that usually embodies the determination of frequencies of the occurrence of specific behaviors, or ratings of their intensity or quality.

Teacher-made achievement tests (objective and essay), the scores on which allow inferences regarding the extent to which specific instructional objectives have been attained.

Teacher-made rating scales and check lists for observation of classroom behaviors: performance levels of speech, music, and art; manifestation of creative endeavors, personal and social adjustment, physical well-being.

Teacher-modified forms (preferably with consultant aid) of the semantic differential scale.

III. Indicators of Status or Change in Student Behaviors Other than Those Measured by Tests, Inventories, and Observation Scales in Relation to the Task of Evaluating Objectives of School Programs

Absences: full-day, half-day, and other selective indices pertaining to frequency and duration of lack of attendance.

*Appended material to paper entitled "Paradigm Involving Multiple Criterion Measures for the Evaluation of the Effectiveness of School Programs" presented at the 1967 Annual Meeting of AERA, February 16, 1967, held in New York City.

Anecdotal records: critical incidents noted including frequencies of behaviors judged to be highly undesirable or highly deserving of commendation.

Appointments: frequencies with which they are kept or broken.

Articles and stories: numbers and types published in school newspapers, magazines, journals, or proceedings of student organizations.

Assignments: numbers and types completed with some sort of quality rating or mark attached.

Attendance: frequency and duration when attendance is required or considered optional (as in club meetings, special events, or off-campus activities).

Autobiographical data: behaviors reported that could be classified and subsequently assigned judgmental values concerning their appropriateness relative to specific objectives concerned with human development.

Awards, citations, honors, and related indicators of distinctive or creative performance: frequency of occurrence of judgments of merit in terms of scaled values.

Books: numbers checked out of library, numbers renewed, numbers reported read when reading is required or when voluntary.

Case histories: critical incidents and other passages reflecting quantifiable categories of behavior.

Changes in program or in teacher as requested by student: frequency of occurrence.

Choices expressed or carried out: vocational, avocational, and educational (especially in relation to their judged appropriateness to known physical, intellectual, emotional, social, aesthetic, interest, and other factors).

Citations: commendatory in both formal and informal media of communication such as in the newspaper, television, school assembly, classroom, bulletin board, or elsewhere (see Awards).

"Contracts": frequency or duration of direct or indirect communications between persons observed and one or more significant others with specific reference to increase or decrease in frequency or to duration relative to selected time intervals.

Disciplinary actions taken: frequency and type.

Dropouts: numbers of students leaving school before completion of program of studies.

Elected positions: numbers and types held in class, student body, or out-of-school social groups.

Extracurricular activities: frequency or duration of participation in observable behaviors amenable to classification such as taking part in athletic events, charity drives, cultural activities, and numerous service-related avocational endeavors.

Grade placement: including numbers of recommended units of course work in academic as well as in non-college preparatory programs.

Grouping: frequency and/or duration of moves from one instructional group to another within a given class grade.

Homework assignments: punctuality of completion, quantifiable judgments of quality such as class marks.

Leisure activities: numbers and types of; times spent in; awards and prizes received in participation.

Library card: possessed or not possesses; renewed or not renewed.

Load: numbers of units or courses carried by students.

Peer group participation: frequency and duration of activity in what are judged to be socially acceptable and socially undesirable behaviors.

Performance: awards, citations received; extra-credit assignments and associated points earned; numbers of books or other learning materials taken out of the library; products exhibited at competitive events.

Performance: awards, citations received; extra-credit assignments and associated points earned; numbers of books or other learning materials taken out of the library; products exhibited at competitive events.

Recommendations: numbers of and judged levels of favorableness.

Recidivism by students: incidents (presence or absence or frequency of occurrence) of a given student's returning to a probationary status, to a detention facility, or to observable behavior patterns judged to be socially undesirable (intoxicated state, dope addiction, hostile acts including arrests, sexual deviation).

Referrals: by teacher to counselor, psychologist, or administrator for disciplinary action, for special aid in overcoming learning difficulties, for behavior disorders, for health defects or for part-time employment activities.

Referrals: by student himself (presence, absence, or frequency).

Service points: numbers earned.

Skills: demonstration of new or increased competencies such as those found in physical education, crafts, homemaking, and the arts that are not measured in a highly valid fashion by available tests and scales.

Social mobility: numbers of times student has moved from one neighborhood to another and/or frequency which parents have changed jobs.

Tape recordings: critical incidents contained and other analyzable events amenable to classification and enumeration.

Tardiness: frequency of.

Transiency: incidents of.

Transfers: numbers of students entering school from another school (horizontal move).

Withdrawal: numbers of students withdrawing from school or from a special program (see Dropouts).

IV. Indicators of Status or Change in Cognitive and Affective Behaviors of Teachers and other School Personnel in Relation to the Evaluation of School Programs.

Articles: frequency and types of articles and written documents prepared by teachers for publication or distribution.

Attendance: frequency of, at professional meetings or at inservice training programs, institutes, summer schools, colleges and universities (for advanced training) from which inferences can be drawn regarding the professional person's desire to improve his competence.

Elective offices: numbers and types of appointments held in professional and social organizations.

Grade point average: earned in postgraduate courses.

Load carried by teacher: teacher-pupil or counselor-pupil ratio.

Mail: frequency of positive and negative statements in written correspondence about teachers, counselors, administrators, and other personnel.

Memberships including elective positions held in professional and community organizations: frequency and duration of association.

Model congruence index: determination of how well the actions of professional personnel in a program approximate certain operationally-stated judgmental criteria concerning the qualities of a meritorious program.

Moonlighting: frequency of outside jobs and time spent in these activities by teachers or other school personnel.

Nominations by peers, students, administrators, or parents for outstanding service and/or professional competencies: frequency of.

Rating scales and check lists (e.g., graphic rating scales or the semantic differential) of operationally-stated dimensions of teachers' behaviors in the classroom or of administrators' behaviors in the school setting from which observers may formulate inferences regarding changes of behavior that reflect what are judged to be desirable gains in professional competence, skills, attitudes, adjustment, interests, and work efficiency; the perceptions of various members of the total school community (parents, teachers, administrators, counselors, students, and classified employees) of the behaviors of other members may also be obtained and compared.

Records and reporting procedures practiced by administrators, counselors and teachers: judgments of adequacy by outside consultants.

Termination: frequency of voluntary or involuntary resignation or dismissals of school personnel.

Transfers: frequency of requests of teachers to move from one school to another.

V. Indicators of Community Behaviors in Relation to the Evaluation of School Programs

Alumni participation: numbers of visitations, extent of involvement in PTA activities, amount of support of a tangible (financial) or a service nature to a continuing school program or activity.

Attendance at special school events, at meetings of the board of education, or at other group activities by parents: frequency of.

Conferences of parent-teacher, parent-counselor, parent-administrator sought by parents: frequency of request.

Conferences of the same type sought and initiated by school personnel: frequency of requests and record of appointments kept by parents.

Interview responses amenable to classification and quantification.

Letters (mail): frequency of requests for information, materials, and servicing.

Letters: frequency of praiseworthy or critical comments about school programs and services and about the personnel participating in them.

Participant analysis of alumni: determination of locale of graduates, occupation, affiliation with particular institutions, or outside agencies.

Parental response to letters and report cards upon written or oral request of school personnel: frequency of compliance by parents.

Telephone calls from parents, alumni, and from personnel in communications media (e.g., newspaper reporters): frequency, duration, and quantifiable judgments about statements monitored from telephone conversations.

Transportation requests: frequency of.

Introductory Note--Cluster 4: "Attitudes, Judgments, and Opinions"

As contemporary events clearly document, evaluation is, above all, a political endeavor. Political action is the basis for implementing reform movements in a complex, democratic society, and the call for increased accountability cannot be isolated from a widely-held desire that education should be changed.

The first of two items in this concluding section deals with controversies generated by performance contracting. It is followed by materials related to analysis of the attitudes we hold about evaluation, for those who assess and ultimately judge are not without their own biases. We cannot escape them; but we can discover more information about our beliefs as well as those of others.

GAIN SCORE ERRORS IN PERFORMANCE CONTRACTING

Robert E. Stake and James L. Wardrop
University of Illinois at Urbana-Champaign

(Editor's Note: The major concerns which underlie the comments appearing in this item focus on the use of individual gain scores as a basis for payment to a performance contractor. The authors' principal reservation is with the erroneous belief (accepted by many schoolboards) that short-term achievement of individual students can be measured reliably by standardized achievement tests. According to their paper, one should expect that **25 percent of the students will show a year's gain in achievement entirely due to the error in such tests.**

According to Stake and Wardrop's figures, if students were tested and retested on a parallel form of the test the next day, one should expect to have one child in four grow "miraculously" a year or more in achievement! Thus, the performance contractor whose basic fee can be covered by payment for one out of four students having "grown a year or more," as measured by a standardized test, may be in a no-risk business, due to the schoolmen's lack of knowledge of standardized test reliability.)

Recent efforts to evaluate the effectiveness of instruction--particularly in performance contracting--reflect confidence that short-term achievement can be measured reliably by standardized achievement tests. Contracts such as the Dorsett-Texarkana contract (Andrew and Roberts, 1970) pay off on an individual-student basis. The contractor typically is to be reimbursed for each student who gains more than a specified amount. The student is to be tested, trained, and retested with a carefully normed, commercially published achievement test. For such tests, and for these only, scores can be reported in terms of the grade equivalent, a publicly interpretable indicator of student academic standing. The contract, for example, might call for termination of remedial training when the student shows a one-year gain in his grade-equivalent score.

Lord (1956) and Webster and Bereiter (1963) have demonstrated that such gain scores are unreliable. At present, this unreliability of gain scores (on two parallel forms of the same test) is such that in reading, for example, we should expect 25% of the students to show a year's gain in achievement merely as an artifact of the error in testing. (Of course, 25% would also show a year's loss.) Nevertheless, with encouragement from the U. S. Office of Education, school districts are contracting with commercial firms for instruction with reimbursement based on individual gain scores. It appears to us that this criterion should be challenged by specialists in educational measurement.

Consider the error in these grade-equivalent scores. For a typical standardized test, the Technical Manual provides such information as the following:

> Reliability of each of two parallel forms equals .84
> Intercorrelation between scores on the two forms equals .81
> Standard deviation for either form (grade equivalents) equals 2.7 years

Using these data, we may apply the conventional formula (Thorndike and Hagen, 1961, p. 192) and find the reliability of the difference between each student's scores on the two forms:

> Reliability of difference scores equals .16

We find the standard deviation of those differences in the usual way (e.g., Glass and Stanley, 1970, p. 128):

> Standard deviation of difference scores equals 1.66 years

Knowing the standard deviation (SD) and the reliability (R) of these difference scores, we can find the standard error (SE) of those differences as SE equals SD 1-R :

> Standard error of the difference equals 1.52 years

The probable error (PE) is about 1.0 years (PE equals .6475SE). On the average, errors for 50% of the students in a group would exceed the probable error. That is, approximately one student in four would show a "gain" (and one in four would show a "loss) of at least one year when tested with one form of the test and then retested with the other form simply as a result of the errors of measurement of the test. Here are two more ways to express this result in the typical performance-contracting situation (we are ignoring the gain that might result simply from exposure to the pretest):

> Suppose that three students were to be tested with a parallel form immediately after the pretest. The chances are better than 50-50 that at least one of the three would have gained a year or more and appear ready to graduate from the program.
>
> Suppose that 100 students were admitted to contract instruction and pretested. After a period of time involving no training, they were tested again and the students "gaining" a year were graduated. After another period of time without training, another test and another graduation occur. After the fourth such "terminal" testing--even though no instruction had occurred--the chances are better than 50-50 that **two-thirds** of the students would have graduated.

Obviously the unreliability of gain scores in such circumstances will assure the appearance of learning even when there is no learning at all.

To reduce the magnitude of these errors would require increasing the reliability or decreasing the standard deviation of the test forms. (Verify for yourself that the intercorrelation between test forms will have no effect on the probable error of the difference scores.) Increasing test reliability offers little hope, for in our example increasing the reliability to .96 would leave us with a probable error of more than one-half year. Nor is it reasonable to reduce the standard deviation by seeking a more homogeneous reference group, for the increased homogeneity would also lower reliability. The problem is not one of reference group; it is a problem of validity. The conventional achievement test does not have the necessary content validity for individual student assessment. For years test authors and test publishers have cautioned users against using these tests as diagnostic instruments. Performance-contract criterion tests should, in effect, be diagnostic tests.

Measurement consultants and the school district's evaluator should insist on a criterion procedure that is valid. (Conventional reliability is not essential; small measurement error is.) Criterion testing might involve the use of specially developed criterion-referenced items, performance simulations, and clinical observations and professional judgments. None of these is currently transformable into grade equivalents. Grade equivalents come from standardized tests. If the advantages of standardized tests (grade equivalents, content selection by experts, technical editing, objective scoring, ready availability, etc.) are desired, the contract should be based on group performance rather than individual-student performance scores.

There are other hazards in measuring and interpreting such scores. Regression effects, inappropriate control groups, unwarrented similarities and dissimilarities in teaching and testing materials, misrepresentation of objectives, and unwarranted extrapolation are some that Stake (1971) and Wardrop (1971) have described.

This brief look at the unreliability of gain scores does not indicate whether performance contracting is an appropriate remedy for a district's instructional weaknesses. Standardized tests continue to be valid for discriminating among students and among districts for various educational purposes. This look at the unreliability of gain scores does indicate that individual-student gain on a currently available standardized test should not be used as a criterion of successful instruction.

EVALUATION EXPECTATIONS

The Director of Research is talking with the Associate Superintendent for Instruction. They are talking about their new programs in reading.

Director of Research: "I think if we switch our end-of-the-year test to an inventory with better content validity, we will find out whether or not we should continue these programs. That change and the addition of two more control groups getting the traditional materials will put us in a much more defensible position insofar as our evaluation plan is concerned."

Associate Superintendent: "That's a good idea, but I know our teachers need help in using the new materials and the parents of our children really don't know what the program is all about. When the parents and teachers understand and support a program, its effectiveness increases considerably."

How is it possible for these two educators to realize that they have different viewpoints about evaluating curricula? What does each of them expect that an evaluation study will do for the reading programs?

One way of examining their different viewpoints is with the help of the CIRCE Attitude Scale 1.3. Through the use of this scale, it is possible to create a profile for both the Associate Superintendent and the Research Director which will enable them to compare themselves and their responses to individual items.

After taking and self-scoring the 48-item inventory, the Assistant Superintendent and the Research Director could each sketch his profile by connecting his responses with a penciled line. You can create their profiles by doing just that. An illustration follows on the next page.

ASSISTANT SUPERINTENDENT

ORIENTATION TO EVALUATION	SCALE SCORES											
Research	0	1	2	[3]	4	5	6	7	8	9	10	11
Service	0	1	2	3	4	5	6	7	8	[9]	10	11
Teaching	0	1	2	3	4	5	6	7	8	[9]	10	11
Objectives	0	1	2	3	[4]	5	6	7	8	9	10	11
Judgment	0	1	2	3	4	5	[6]	7	8	9	10	11
Confidence (in evaluation)	0	1	2	3	4	5	[6]	7	8	9	10	11

RESEARCH DIRECTOR

ORIENTATION TO EVALUATION	SCALE SCORES											
Research	0	1	2	3	4	5	6	7	8	[9]	10	11
Service	0	1	2	3	4	5	[6]	7	8	9	10	11
Teaching	0	1	2	3	4	[5]	6	7	8	9	10	11
Objectives	0	1	2	3	4	5	6	7	8	[9]	10	11
Judgment	0	1	2	3	[4]	5	6	7	8	9	10	11
Confidence (in evaluation)	0	1	2	3	4	[5]	6	7	8	9	10	11

The CIRCE Attitude Scale 1.3 is not yet fully developed but early tryouts indicate that it has some validity and that it helps people talk about their viewpoints toward remedying this failure to communicate.

CIRCE Attitude Scale No. 1.4 Name _____

Attitudes toward Educational Evaluation. Below are a number of statements about the evaluation of educational programs. A program can be a lesson, a course, a whole curriculum, or any training activity. Consider each statement as a statement of opinion. If you agree at least a little bit with the statement, Circle the letter A. If you disagree even a little bit with the statement, circle the letter D. If you both agree and disagree, or if you have no opinion, leave the letters uncircled.

A = AGREE D = DISAGREE Blank = Neither

1.	A	D	The major purpose of an educational evaluation study should be to gather information that will be helpful to the educators.
2.	A	D	It is important for the program evaluator to find out how well various people like the program.
3.	A	D	Generally speaking, an educational program should be evaluated with reference to one or more "control" programs.
4.	A	D	The evaluator should accept the responsibility of finding the strongest, most defensible, and publicly attractive points of the program.
5.	A	D	In evaluating a program, it is at least as important to study and report on the types of teaching as it is to study and report on the amount of learning.
6.	A	D	The evaluator should draw a conclusion as to whether or not the goals of the program are worthwhile.
7.	A	D	It is more important to evaluate a program in comparison to what other programs do than to evaluate it with reference to what its objectives say it should do.
8.	A	D	Principals and superintendents should not gather data about the quality of instruction in the classroom.
9.	A	D	The task of putting educational objectives into writing is more the responsibility of the evaluator than that of the educator.
10.	A	D	It is essential that the full array of educational objectives be stated before the program begins.
11.	A	D	Evaluation studies would improve if they gathered more kinds of information, even if at the expense of gathering less reliable information.
12.	A	D	Evaluators should ignore data that cannot be objectively verified.
13.	A	D	Education should have more of an engineering orientation than it now has.
14.	A	D	The job of an evaluator is mostly one of finding out how well students learn what they are supposed to learn.
15.	A	D	Evaluation should aid an educator in revising his goals even while the program is in progress.
16.	A	D	The process of decision-making about the curriculum is one of the weakest links in the present operation of the schools.
17.	A	D	Educators have some important aims that cannot be stated adequately by anyone in terms of student behaviors.
18.	A	D	Information from an evaluation study is not worth the trouble it makes.
19.	A	D	The first job in instruction is the formulation of a statement of objectives.
20.	A	D	A teacher should tell his students any and all of his teaching objectives.
21.	A	D	The major purpose of educational evaluation is to find out the worth of what is happening.
22.	A	D	The evaluator should be a facilitator more than a critic or reformer or scholar.
23.	A	D	Some school experiences are desirable because they round out a child's life-whether or not they increase his competence or change his attitudes.

24. A D An evaluator should find out if the teaching is in fact the kind that the school faculty expects it to be.

25. A D Whether or not an evaluation report is any good should be decided pretty much on the same grounds that research journal editors use to decide whether or not a manuscript should be published.

26. A D The **main** purpose of evaluation is to gain understanding of the causes of good instruction.

27. A D Description and value judgment are equally important components of evaluation.

28. A D In conducting an evaluation, there is no justification for the exercise of subjective judgment of any kind by the evaluator.

29. A D Educational evaluation is a necessary step in the everyday operation of the school.

30. A D The strategy of evaluation should be chosen primarily in terms of the particular needs the sponsors have for evaluation data.

31. A D The educational evaluator should attempt to conceal all of his personal judgment of the worth of the program he is evaluating.

32. A D The sponsor of an evaluation should have the final say-so in choosing or eliminating variables to be studied.

33. A D The main purpose of educational evaluation is to find out what methods of instruction work for different learning situations.

34. A D Parents' attitudes should be measured as part of the evaluation of school programs.

35. A D An evaluator finds it almost impossible to do his job without intruding upon the operation of the program at least a little.

36. A D All important educational aims can be expressed in terms of student behaviors.

37. A D Some educational goals are best expressed in terms of teacher behaviors.

38. A D It is essential that evaluation studies be designed so that the findings are generalizable to other curricula.

39. A D An evaluation study should pay less attention to the statistical significance of a finding than an instructional research study would.

40. A D Evaluation interferes with the running of schools more than it helps.

41. A D Little evaluation planning can be done before you get a statement of instructional objectives.

42. A D The leader of an evaluation team should be a teacher.

43. A D The entire school day and the entire school experience should be divided up and assigned to the pursuit of stated educational goals.

44. A D An evaluation of an educational program should **include** a critical analysis of the value of the goals of the program.

45. A D Every teacher should have formal ways of gathering information about the strengths and shortcomings of his instructional program.

46. A D Money spent on evaluation contributes more to the improvement of education than any other expenditure.

47. A D There just is no way that careful and honest evaluation can hurt a school program.

48. A D If an evaluation study is well designed, the primary findings are likely to improve decisions made by administrators, teachers, and students themselves.

49. A D When the evaluator has to choose between helping this staff run its program better and helping educators everywhere understand all programs a little better he should choose the latter.

Linking Section 5 -- An Interpretive Note

The reluctance of educators to view evaluation as a reliable aid in decision-making is often reinforced by the esoteric terminology which they see as synonomous with the field.

The Glossary of Terms in Part VI is not an exhaustive compilation but rather is a selection of pertinent terms taken from a much larger source. It should assist readers of the Resource Unit in making the widest possible use of materials found here. That span of utility includes attempts to anticipate future problems in education. For example, the current interest in performance or behavioral objectives is likely to result in decreased emphasis on content coverage, an outcome which has received little attention to date.

PART VI--A GLOSSARY OF TERMS

ACCOUNTABILITY

In evaluation, the process during which one provides relevant audiences with descriptions and/or explanations for which one is responsible.

ACHIEVEMENT AGE

The age for which a given score on an achievement test is the real or estimated age. Also called educational age or subject age.

ACHIEVEMENT TEST

An ability test designed to assess the amount an individual has learned in a specified subject area as a result of past experience or training.

AFFECTIVE DOMAIN

One of three major categories for classifying educational goals. According to Krathwohl, Bloom, and Masia the affective domain consists of "objectives which emphasize a feeling tone, an emotion, or a degree of acceptance or rejection. Affective objectives vary from simple attention to complex but internally consistent qualities of character and conscience." Included in the taxonomy of the affective domain are the following five major categories: (1) Receiving (Attending), (2) Responding, (3) Valuing, (4) Organization, and (5) Characterization by a Value or Value Complex. Sub-categories are also available.

AGE NORMS

Values representing the chronological age at which a given level of behavioral development is normally, or on the average, attained and to which test scores are sometimes converted. Essentially a norming system based on age equivalents.

ANECDOTAL RECORD

Usually, a written account describing an observation of an event or an incident of an individual's behavior. Anecodotes may also be transcribed on tape recorders or by other audio-visual means. The anecdote typically should contain a description of what the individual did and said as well as a description of the situation in which the behavior occurred. Under usual conditions one would record incidents which appear relevant for an understanding of the individual, either as being atypical or usual.

APTITUDE

The capacity or extent to which an individual may be expected to acquire a particular kind of ability. It also may be defined as an ability to learn or a readiness for learning.

APTITUDE TEST

An ability test designed to assess or, more precisely, to permit predictions of what the individual, under standard conditions, can learn to do. Aptitude tests may be placed in one of two categories: measures of general scholastic aptitude (the general intelligence test) and measures of specific aptitudes (music, art, foreign language, etc.)

ASSESSMENT

Any of a number of procedures for making relevant evaluations or differentiations among individuals or groups in respect to any characteristic, attribute, or product.

BEHAVIORAL OBJECTIVES

Educational goals which are stated in observable terms and reflect what the student will be like or will be able to do after instruction.

---

*Adapted from **Referenced Glossary of Terms in Evaluation and Measurement,** Alan Ross Coller, ERIC Clearinghouse on Early Childhood Education, University of Illinois, Champaign-Urbana, Illinois 61801.

CHECKLISTS

Instruments specifically designed to collect and record relatively unrefined judgments systematically. An observer indicates, usually by use of checkmarks on a printed form, whether or not the person, place, thing, or event--real or abstract--has the characteristic as specified by the **checklist** item. Data from checklists are recorded as if from a yes or no form and should be contrasted against data from the more refined **rating scales** which are in the form of scores representing points along a **continuum.**

CHRONOLOGICAL AGE (CA)

The real-life age of an organism as expressed in some unit of time; the time elapsed since birth. In measurement, chronological age (CA) is most conveniently expressed in units of months which often calls for the individual's real age to be modified to fit the requirements of the particular test. For example, a child who is 5 years, two months, and 17 days may be assigned the CA of 62 months in one test and 63 months in another test.

COGNITIVE DOMAIN

One of the three major categories for classifying educational **goals.** According to Bloom and his colleagues "the cognitive domain. . .includes those objectives which deal with the recall or recognition of knowledge and the development of intellectual abilities and skills." Included in the taxonomy of the cognitive domain are the following six categories: (1) Knowledge, (2) Comprehension, (3) Application, (4) Analysis, (5) Synthesis, and (6) Evaluation.

CRITERION BASED STANDARDS

An absolute type of standard derived from the **mastery** performance desired on any given task. To be distinguished from **empirical** norms and **estimated** norms.

CRITERION-REFERENCED TESTS

Any test the interpretation of which is concerned with estimating the degree to which an individual's achievements have progressed toward some **criterion** or **standard.** Such tests, also called **mastery tests** (because they are used to evaluate **mastery learning**) are not intended to provide scores that will rank students in terms of their achievements (norm or individually referenced tests do this); rather they provide "absolute" scores employed to classify students as belonging to either one of at least two groups--those whose achievements resemble criterion achievement, and those whose achievements do not. These tests are frequently used in conjunction with programmed instruction since these tests make specific learning difficulties relatively easy to diagnose.

CRITERION-RELATED VALIDITY

A general term used to refer to those measures of **validity** which are employed to determine the degree of association between performances on those tests designed either to forecast an individual's future standing or current standing on a variable different from the tests and between the variable itself. If the test is to be used to forecast the individual's future standings on some dimension we are concerned with **predictive validity.** If the test is to be used to determine the individual's current standing in respect to some dimension, we are concerned with **concurrent validity. Criterion-related validity** is also called **empirical validity.**

DESCRIPTIVE STATISTICS

Those methods employed to **describe** the characteristics of all those members of a group for which data is available. Descriptive statistics may be employed to describe populations (e.g. the U.S. Census employs such statistics to count and categorize the population) and samples drawn from populations. However, in the latter instance, the sample is regarded as the group not the representative population from which the sample was drawn. **Inferential statistics** are employed to make inferences about some larger group on the basis of samples.

DIAGNOSTIC EVALUATION

A type of evaluation designed either for purposes of student placement or for determining the underlying causes for learning deficiencies in the student. When **diagnostic evaluation** is performed prior to instruction it is usually intended for student placement; when performed during instruction its main function is the discovery of the causes for learning difficulties.

DIAGNOSTIC TEST

A type of test designed to identify problem areas. Diagnostic achievement tests are used to pinpoint specific areas of subject-matter strength and weakness, and to determine the exact nature of the strengths and weaknesses. When strengths and weaknesses have been diagnosed, remedial action may be instituted. Diagnostic tests may be contrasted with **survey tests** which are designed to provide only a general overview of subject-matter competence.

EDUCATIONAL MEASUREMENT

Refers to the employment of data-gathering techniques in order to collect precise and more or less objective descriptive information about the educational process. The more parochial view regards measurement as only a part, although an important part, of the total process of evaluation. This latter view treats the results of measurement as numbers which describe or express something about the characteristics observed and evaluation as the process by which judgments about these descriptive numbers are made.

EDUCATIONAL OBJECTIVES

A statement of instructional purpose which describes the ways in which students are to be changed by their interaction with the educational situation. Educational objectives would communicate the educator's intent and be expressed in terms of observable student behavior and content areas. It also has been suggested that **educational objectives** should as well contain a description of how the student is to demonstrate that he has achieved the objective and the important conditions under which the behavior is expected to occur and the accuracy of the anticipated performance.

EMPIRICAL

Based upon a factual investigation or experience rather than upon reason or theory. An **empirical** evaluation, for example, is one in which one might try to find out how much students actually learned from various textbooks. A non-empirical evaluation might be composed of a group of teachers analyzing the contents of several texts to determine from which of them the students would best learn.

EMPIRICAL NORMS

In measurement, standards based upon the responses of a representative sample of examinees of the type for whom the test was constructed. Such standards, also called **norms**, are likely to occur with published tests. To be distinguished from estimated norms and **criteria-based performance.**

FACE VALIDITY

A judgment of whether a test or technique measures what it looks like it should measure. The cooperativeness of the examinee may be affected by the degree to which test items appear to be related to the stated aims of the test. If the test items seem unreasonable the examinee may not be adequately motivated to do his best. Unlike other estimates of **validity**, **face validity** does not ordinarily result in a numerical index but rather a qualitative one.

FACTOR ANALYSIS

Any of a set of procedures employed to analyze the interrelationships among a set of variables. By a process which assesses the proportion of **variance** of each variable that is associated with a limited set of factors; an assumption of commonality that different variables have many aspects in common; that they may be measuring the same things, etc.

FORMATIVE EVALUATION

Systematic evaluation that occurs before the terminating point of a segment of instruction, that is, during the learning process when the student is undergoing change. Such evaluations are useful for curriculum construction, teaching, and learning. For example, **formative evaluation** procedures may be employed to indicate areas in which remediation is needed so that instruction can be so modified. Formative evaluation is distinguished from summative evaluation in that it occurs **during** the operational sequence of a program or project. "Feedback" is part of formative evaluation.

GENERAL INTELLIGENCE TEST

A general aptitude type of ability test, sometimes called a **general ability test**, employed in the prediction of **scholastic ability**. Existing tests are thought to be composed of both **crystallized** and **fluid general ability** items.

INFERENTIAL STATISTICS

Those methods employed to make estimates or inferences concerning the characteristics of a larger group on the basis of only a **sample** from that group. **Inferential statistics** is to be distinguished from **descriptive statistics** which is primarily concerned with describing the characteristics of a group.

NORM-REFERENCED TESTS

Any test, the interpretation of which is primarily concerned with estimating the level of an individual's achievements in respect to other individuals--the norming population. The individual's "relative" standing, as contrasted to the "absolute" standard employed in **criterion-referenced tests**, is the essential purpose of the measurement. Such tests, also called **individually-referenced tests**, are usually comprised of items adjusted to a 50-60% item difficulty level. In contrast, criterion-referenced tests usually have item difficulty levels of about 85%.

NORMS

Typically, represent the average performance level on tests or in activities with which individuals or groups may be compared. In the school situation, norms are usually employed to indicate average performances vis-a-vis different age and grade levels. In evaluation, the comparison of local observations with norms forms the basis of the judgmental process.

PERFORMANCE CHECKLIST

A type of checklist designed especially to collect and record relatively unrefined judgments concerned with the behavior or performance of individuals or groups. Such checklists, also called behavior checklists, require the observer to simply indicate, usually by checkmarks, whether the behavior of the individual or group concerned has the characteristics as specified by the checklist item.

BEHAVIOR RATING SCALE

A type of rating scale designed especially to collect and record systematic and refined judgments concerned with the behavior (or performance, or characteristics, or traits) of individuals or, in some cases, groups. While the basic form of behavior rating scales may vary to a considerable extent, all such instruments employ underlying behavioral continuums and rely upon a rater to make observations of the individual being rated. The behavior in question may be cognitive, affective, or psychomotor.

SITUATION TEST

A procedure whereby the behavior of all individuals (or small group of persons) is observed while performing in a more or less structured situation. Some situations call for role playing as in the psychodrama technique or the more modern encounter groups. Other situations may call for discussions, problem solving, or cooperative ventures.

SUMMATIVE EVALUATION

Systematic evaluation that takes place after the termination of an instructional segment; i.e., at the end of a unit chapter, course, or semester. Summative evaluation generally requires judgments; its primary goals being grading, selecting, certifying, providing feedback, judging teacher effectiveness, and comparing curricula. An intermediate-type of summative evaluation is to be distinguished from a longer-term evaluation. Of the two, the former is more concerned with outcomes that are more direct, less generalizable and less transferable. Summative evaluation is also to be distinguished from formative evaluation which takes place before the termination of an instructional segment.

TASK ANALYSIS

The process by which a desired outcome is described and then analyzed into behavioral components which must be sequentially developed in order to arrive at the terminal behavior--the desired outcome. A prime purpose of task analysis is to apply findings derived from learning theory to the instructional sequence--the set of hierarchical steps.